Постановка задачи

1. Получить данные временного ряда
2. Оценить временной ряд на стационарность двумя способами
    a. Провести визуальную оценку ряда и скользящей статистики
    b. Тест Дики-Фуллера
3. Разложить временной ряд на тренд, сезональность и остаток в соответствии с аддитивной и мультипликативной моделями
4. Проанализировать стационарность полученных рядов
5. Проверить, является ли ряд интегрированным порядка k. Если да, то применить модель ARIMA.

Временной ряд

Под **временным рядом** понимаются последовательно измеренные через некоторые (зачастую равные) промежутки времени данные. **Прогнозирование временных рядов** заключается в построении модели для предсказания будущих событий основываясь на известных событий прошлого, предсказания будущих данных до того как они будут измерены.

Стационарность

Математическое ожидание – среднее значение случайной величины при стремлении количества выборок или количества испытаний к бесконечности.

Дисперсия – мера разброса значений случайно величины относительно ее математического ожидания.

Автоковариационная функция – совокупность значений ковариаций при всевозможных значениях расстояния между моментами времени.

Ряд называется **строго стационарным** или **стационарным в узком смысле**, если сдвиг во времени не меняет ни одну из функций плотности распределения.

Ряд называется **слабо стационарным** или **стационарным в широком смысле**, если такие статистические характеристики временного ряда как его математическое ожидание и дисперсия существуют и не зависят от времени, а автокорреляционная (автоковариационная) функция зависит только от величины $(t_1 – t_2)$.

Если нарушается хотя бы одно из этих условий, то ряд называется нестационарным.

Скользящая статистика

Скользящее среднее (Moving Average, MA)– среднее арифметическое значений исходной функции за установленный период. Позволяет выявить основные тенденции, сглаживая краткосрочные колебания.

Стандартное отклонение (Standart Deviation, STD) – показывает, на сколько в среднем отклонится ряд от среднего арифметического.

Ковариация - в теории вероятностей и математической статистике мера линейной зависимости двух случайных величин

Интегрированность

Ряд называется интегрированным порядка k ($y_t \sim I(k)$), если разности ряда k-го порядка являются стационарными, в то время как разности меньшего порядка не являются стационарными.

Тест Дики-Фуллера

Тест Дики-Фуллера является средством проверки стационарности временного ряда. Является одним из тестов на единичные корни.

Авторегрессионная модель (AR-модель) – временных рядом, в которой значения временного ряда в данный момент линейно зависят от предыдущих его значений.

Авторегрессионный процесс порядка p (AR(p)-процесс) – прцоесс, который определяется следующим образом:

$$y_t = c + \sum_{i=1}^{p} a_i y_{t-i} + \varepsilon_{t,}$$

Где $a_1,..,a_p$ – параметры модели (коэффициенты авторегрессии), c -константа, ε - белый шум.

Стационарность авторегрессионого процесса зависит от корней характеристического полинома

$$a(z) = 1 - \sum_{i=1}^{n} a_i z_i,$$

Для того, чтобы процесс был стационарным, достаточно потребовать $|z| > 1$.

Если характеристическое уравнение $a(z) = 0$ имеет корни, равные по модулю единицу, то эти корни называются **единичными**.
Временной ряд имеет хотя бы один **единичный корень** или порядок интеграции один, если его первые разности образуют стационарный ряд. y(t) ~ I(1)

Тест Дики-Фуллера проверяет значение коэффициента a в авторегрессионом уравнении первого порядка, вида:

$$y_t = a y_{t-1} + \varepsilon_t,$$

Результат:

- $a = 1$ – *есть единичные корни, стационарности нет*
- $|a| < 1$ – нет единичных корней, есть стационарность
- $|a| > 1$ – не свойственно для временных рядов, которые встречаются в реальной жизни, требуется более сложный анализ

Временные ряды состоят из трех компонент:

1. **Тренд** (T) – изменение, определяющее направление изменений показателей ряда
2. **Сезонность** (S) – повторяющийся краткосрочный цикл в ряде
3. **Остаток** (E) – разница между предсказанным и наблюдаемым значением

Модель, в которой временной ряд представлен в виде суммы этих компонент называется **аддитивной**, в виде произведения компонент – **мультипликативной.**

Центрированное скользящее среднее – скользящее среднее (MA), вычисленное от скользящего среднего.

Аддитивная Y = T + S + E

Сезональность (S) это разность временного ряда и центрированного скользящего среднего.

Тренд (T) находится с помощью метода наименьших квадратов, который приближает временной ряд.
Остаток (E) это временной ряд минус тренд и сезональность.

Мультипликативная $Y = T \times S \times E$

Сезональность (S) в данном случае есть частное временного ряда и центрированного скользящего среднего.
Тренд (T) находится аналогично аддитивной модели.
Остаток (E) это частное временного ряда и частного сезональности и тренда.

Модель ARIMA

Модель скользящего среднего (MA-модель) – в ней моделируемый уровень временного ряда можно представить как линейную функцию прошлых ошибок, то есть разностей между фактическим и теоретическим уровнем.

Авторегрессионная модель скользящего среднего (ARMA-модель) – математическая модель, используемая для анализа и прогнозирования стационарных временных рядов.

Авторегрессионная интегрированная модель скользящего среднего (ARIMA-модель) – модель анализа временных рядов. Является расширением ARMA-модели для нестационарных временных рядов, которые можно сделать стационарными взятием разностей некоторого порядка от исходного временного ряда.

Информационный критерий – применяемая в статистике мера относительного качества статистических моделей, учитывающая степень подгонки модели под данные с корректировкой на используемое количество оцениваемых параметров. Информационные критерии используются исключительно для сравнения моделей между собой, без содержательной интерпретации значений этих критериев. Обычно, чем меньше значение критерия, тем выше относительное качество модели.

Информационный критерий Акаике – критерий, применяющийся только для выбора из нескольких статистических моделей.

Программная реализация

Используемые библиотеки

pandas – библиотека для обработки и анализа данных, предоставляет инструменты для работы с временными рядами
matplotlib – библиотека для визуализации данных
statsmodels – библиотека для проверки различных статистических гипотез
statsmodels.tsa – библиотека, упрощающая работы с временными рядами
sklearn – библиотека, содержащая методы регрессионного анализа и алгоритмы машинного обучения

Используемые функции

diff(k) – взятие разности ряда порядка k

iloc – взятие серии от дата фрейма в библиотеке pandas

adfuller из библиотеки *statsmodels.tsa* – тест Дики_фуллера

read_excel из библиотеки *pandas* – считывает данные из excel файла

rolling.mean() – вычисление скользящего среднего (MA)

rolling.std() – вычисление стандартного отклонения (STD)

seasonal-decompose – разложение временного ряда на тренд, сезональность и остаток в зависимости от модели (аддитивной или мультипликативной)

Реализованные функции

Integ – определяет порядок интегрирования

DFuller – проводит тест Дики-Фуллера и выводит результат

Работу выполнили

Ворончихина Анастасия – Написание кода
Тангаев Артем – Написание readme